

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平10-187182

(43)公開日 平成10年(1998) 7月14日

(51)Int.Cl.⁶
G 1 0 L 3/00

H 0 4 N 5/91

識別記号
5 3 1
5 1 5
5 5 1

F I
G 1 0 L 3/00

H 0 4 N 5/91

5 3 1 N
5 1 5 B
5 5 1 G
N

審査請求 未請求 請求項の数12 O L (全 8 頁)

(21)出願番号 特願平8-340293

(22)出願日 平成 8 年(1996)12月20日

(71)出願人 000004226

日本電信電話株式会社
東京都新宿区西新宿三丁目19番2号

(72)発明者 南 憲一

東京都新宿区西新宿3丁目19番2号 日本
電信電話株式会社内

(72)発明者 阿久津 明人

東京都新宿区西新宿3丁目19番2号 日本
電信電話株式会社内

(72)発明者 外村 佳伸

東京都新宿区西新宿3丁目19番2号 日本
電信電話株式会社内

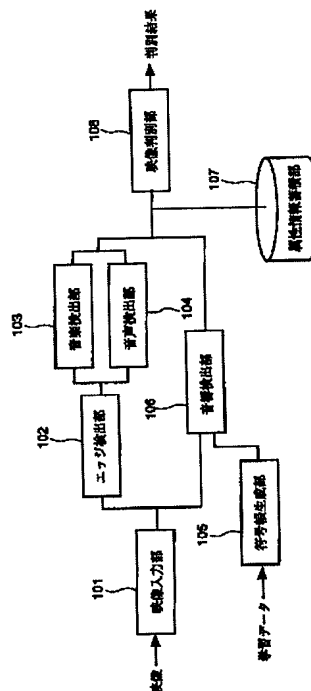
(74)代理人 弁理士 志賀 富士弥

(54)【発明の名称】 映像分類方法および装置

(57)【要約】

【課題】 映像情報に含まれる音情報を解析し、映像を既存のジャンルにとらわれないカテゴリーに分類する映像分類方法および装置を提供する。

【解決手段】 音楽検出部103は、入力された映像情報の音情報を周波数解析し、スペクトルの安定性を検出して音楽を検出する。音声検出部104は、スペクトルのハーモニック構造を検出し、音声を検出する。他方で、入力された映像情報の音情報を、符号帳生成部105にて学習データとして生成された符号帳の特徴ベクトルと音響検出部106において比較し、両者の距離の近さにより音響の種類を検出する。以上で検出された音情報の区間の位置を属性情報蓄積部107で記録し、検出された音情報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置のパターンを抽出して、映像情報の種類を映像判別部108にて判別する。



【特許請求の範囲】

【請求項1】 映像情報を入力し、該入力された映像情報に含まれる音情報から音楽、音声、音響のうち少なくとも1つが存在する区間を検出し、該検出された区間の発生パターンによって映像の種類を判別する映像分類方法であって、

映像情報がアナログの場合にはA/D変換してデジタルの映像情報を入力する映像入力段階と、
該映像情報に含まれる音情報を周波数解析し、スペクトルの安定性を出すエッジ検出段階と、
該スペクトルの安定性から音楽を検出する音楽検出段階と、
該スペクトルのハーモニック構造を検出し、音声を検出する音声検出段階と、
音響の特徴ベクトルを学習データとしてベクトル量子化し、符号帳を生成する符号帳生成段階と、
該生成された符号帳と該映像情報に含まれる音情報の特徴ベクトルとを比較し、距離の近い音響を検出する音響検出段階と、
該検出された音情報の種類別の区間の位置を記録する属性情報蓄積段階と、
該検出された音情報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置のパターンの一以上を抽出し、該映像情報の種類を判別する映像判別段階と、
を有することを特徴とする映像分類方法。

【請求項2】 前記エッジ検出段階では、前記スペクトルを時間方向に並べたスペクトログラムから、周波数方向の微分オペレータによってエッジを検出すること、を特徴とする請求項1に記載の映像分類方法。

【請求項3】 前記音楽検出段階では、前記スペクトログラムの一定周波数における時間方向のエッジの強さから音楽を検出すること、を特徴とする請求項2に記載の映像分類方法。

【請求項4】 前記音声検出段階では、前記スペクトログラムのエッジの強い部分を除去した後に、くし形フィルタを用いてハーモニック構造を検出し、音声を検出すること、を特徴とする請求項2または3に記載の映像分類方法。

【請求項5】 前記音響検出段階では、参照音として一種類の音響のみを含む音情報の特徴ベクトルと、前記符号帳の重心との距離を算出し、距離が最も近くなる頻度の高い該符号帳の重心と、前記映像情報に含まれる音情報の特徴ベクトルとの距離を検出の判定基準として用いること、を特徴とする請求項1、2、3、4のいずれかに記載の映像分類方法。

【請求項6】 前記映像判別段階は、検出された音情報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置を分類ベクトルとして符号帳を作成し、該

符号帳の重心と、前記映像情報に含まれる音情報の分類ベクトルとの距離を判別基準に用いること、を特徴とする請求項1、2、3、4、5のいずれかに記載の映像分類方法。

【請求項7】 映像情報を入力し、該入力された映像情報に含まれる音情報から音楽、音声、音響のうち少なくとも1つが存在する区間を検出し、該検出された区間の発生パターンによって映像の種類を判別する映像分類装置であって、
映像情報がアナログの場合にはA/D変換してデジタルの映像情報を入力する映像入力部と、
該映像情報に含まれる音情報を周波数解析し、スペクトルの安定性を出すエッジ検出部と、
該スペクトルの安定性から音楽を検出する音楽検出部と、
該スペクトルのハーモニック構造を検出し、音声を検出する音声検出部と、
音響の特徴ベクトルを学習データとしてベクトル量子化し、符号帳を生成する符号帳生成部と、
該生成された符号帳と該映像情報に含まれる音情報の特徴ベクトルとを比較し、距離の近い音響を検出する音響検出部と、
該検出された音情報別の区間の位置を記録する属性情報蓄積部と、
該検出された音情報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置のパターンを一以上抽出し、該映像情報の種類を判別する映像判別部と、
を備えることを特徴とする映像分類装置。

【請求項8】 前記エッジ検出部は、前記スペクトルを時間方向に並べたスペクトログラムから、周波数方向の微分オペレータによってエッジを検出するものであること、を特徴とする請求項7に記載の映像分類装置。

【請求項9】 前記音楽検出部は、前記スペクトログラムの一定周波数における時間方向のエッジの強さから音楽を検出するものであること、を特徴とする請求項8に記載の映像分類装置。

【請求項10】 前記音声検出部は、前記スペクトログラムのエッジの強い部分を除去した後に、くし形フィルタを用いてハーモニック構造を検出し、音声を検出するものであること、を特徴とする請求項8または9に記載の映像分類装置。

【請求項11】 前記音響検出部は、参照音として一種類の音響のみを含む音情報の特徴ベクトルと、前記符号帳の重心との距離を算出し、距離が最も近くなる頻度の高い該符号帳の重心と、前記映像情報に含まれる音情報の特徴ベクトルとの距離を検出の判定基準として用いること、を特徴とする請求項7、8、9、10のいずれかに記載の映像分類装置。

【請求項12】 前記映像判別部は、検出された音情報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置を分類ベクトルとして符号帳を作成し、該符号帳の重心と、前記映像情報に含まれる音情報の分類ベクトルとの距離を判別基準に用いるものである、ことを特徴とする請求項7、8、9、10、11のいずれかに記載の映像分類装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 映像を効率良く扱うためには、映像の属性情報を自動的に付与する技術が必要である。属性情報は、映像制作の関連分野において、映像の編集、加工、分類等に利用される。本発明は、映像に含まれる特徴量を抽出し、特徴量に応じて映像を分類する技術に関する。

【0002】

【従来の技術】 映像の内容がどのようなものであるかを大別することは、ビデオ・オン・デマンドのようなシステムで用いられる大量の映像を効率良く扱う上で不可欠である。現在、映像は主にニュース、スポーツ、ドラマ、映画、音楽、ドキュメンタリー、教育、バラエティ、アニメ等に分類されているが、これらのうち幾つかを自動的に識別しようとする方法が提案されている。

「S. Fischer et. al: Automatic Recognition of Film Genres, ACM Multimedia'95, pp. 295-301」では、画像の色情報から場面の変わり目やカメラの動きを検出し、音情報の振幅の変化と併せて、ニュース、スポーツ（テニスおよび自動車レース）、アニメ、コマーシャルの分類を行っている。カメラの動きが少なければニュース、周期的な音の繰り返し（テニスのボールを打つ音）があればスポーツ、言葉が途切れた所にノイズが少なければアニメ（アフレコのため背景音が少ない）、場面の変わり目に全体が黒になればコマーシャルといったようにジャンル毎にみられる典型的な特徴を利用している。

【0003】

【発明が解決しようとする課題】 上記従来の技術では、主に画像情報に基づいて映像の分類を行っており、音情報についての詳しい解析は行われていない。また、画像情報から検出できる、ジャンル毎に固有の特徴が限られているため、分類できる範囲は狭い。さらに、上記のように従来から定められているジャンル毎の特徴を見つけ出すようなトップダウン的な方法では、分類できないジャンルが存在する。

【0004】 一方、映像に含まれる音情報は映像の内容を良く反映しており、内容の種類に固有の特徴を検出し易い。音情報を解析して映像一般に見られる特徴的な音を検出し、その発生パターンから映像を分類することで、ボトムアップ的な要素を取り入れた分類方法を実現

することが可能である。

【0005】 本発明の目的は、映像情報に含まれる音情報を解析し、映像を既存のジャンルにとらわれないカテゴリに分類する映像分類方法および装置を提供することにある。

【0006】

【課題を解決するための手段】 上記の目的を達成するため、本発明の映像分類方法は、映像情報がアナログの場合にはA/D変換してデジタルの映像情報を入力する映像入力段階と、該映像情報に含まれる音情報を周波数解析し、スペクトルの安定性を検出し、音楽を検出する音楽検出段階と、該スペクトルのハーモニック構造を検出し、音声を検出する音声検出段階と、音響の特徴ベクトルを学習データとしてベクトル量子化し、符号帳を生成する符号帳生成段階と、生成された符号帳と該映像情報に含まれる音情報の特徴ベクトルを比較し、距離の近い音響を検出する音響検出段階と、該検出された音情報の種類別の区間の位置を記録する属性情報蓄積段階と、該検出された音情報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置のパターンを一以上抽出し、該映像情報の種類を判別する映像判別段階と、を有することで、入力された映像情報に含まれる音情報から音楽、音声、音響のうち少なくとも1つが存在する区間を検出し、該検出された区間の発生パターンによって映像の種類を判別して広範囲なカテゴリに分類することが可能となる。

【0007】 また、本発明の映像分類装置は、映像情報がアナログの場合にはA/D変換してデジタルの映像情報を入力する映像入力部と、該映像情報に含まれる音情報を周波数解析し、スペクトルの安定性を検出し、音楽を検出する音楽検出部と、該スペクトルのハーモニック構造を検出し、音声を検出する音声検出部と、音響の特徴ベクトルを学習データとしてベクトル量子化し、符号帳を生成する符号帳生成部と、生成された符号帳と該映像情報に含まれる音情報の特徴ベクトルを比較し、距離の近い音響を検出する音響検出部と、該検出された音情報の種類別の区間の位置を記録する属性情報蓄積部と、該検出された音情報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置のパターンを一以上抽出し、該映像情報の種類を判別する映像判別部と、を具備することで、入力された映像情報に含まれる音情報から音楽、音声、音響のうち少なくとも1つが存在する区間を検出し、該検出された区間の発生パターンによって映像の種類を判別して広範囲なカテゴリに分類することが可能となる。

【0008】 上記の映像分類方法および装置では、スペクトログラムの一定周波数における時間方向のエッジの強さを検出することで、音楽を容易に検出することが可能となる。

【0009】 また、該スペクトログラムのエッジの強い

部分を除去した後に、くし形フィルタを用いてハーモニク構造を検出することで、音楽が重なっている場合でも音声を容易に検出することが可能となる。

【0010】また、参照音として一種類の音響のみを含む音情報の特徴ベクトルと、該符号帳の重心との距離を算出し、距離が最も近くなる頻度の高い該符号帳の重心と、該映像情報に含まれる音情報の特徴ベクトルとの距離を検出の判定基準として用いることで、学習した音響を容易に検出することが可能となる。

【0011】さらに、検出された音情報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置を分類ベクトルとして符号帳を作成し、該符号帳の重心と、該映像情報に含まれる音情報の分類ベクトルとの距離を判別基準に用いることで、映像を容易に分類することが可能となる。

【0012】

【発明の実施の形態】次に、本発明の実施の形態について図面を参照して詳細に説明する。

【0013】図1は、本発明の一実施形態例の映像分類装置の概略構成を示すブロック図である。

【0014】本実施形態例の映像分類装置は、映像情報がアナログの場合にはA/D変換して入力する映像入力部101と、音情報を周波数解析して、サウンドスペクトログラムのエッジを検出し、必要に応じて除去するエッジ検出部102と、音情報から音楽を検出する音楽検出部103と、音声を検出する音声検出部104と、音響の学習データから符号帳を生成する符号帳生成部105と、学習した音響と同一種類の音を検出する音響検出部106と、検出された音の区間の位置を記録する属性情報蓄積部107と、検出された音情報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置によって、映像情報の種類を判別する映像判別部108から構成されている。

【0015】映像入力部101から入力された映像の音データは、一方でエッジ検出部102に入力され、エッジ検出部102でFFT（高速フーリエ変換）処理されて、数秒程度の長さのサウンドスペクトログラムが生成される。ここで、FFTの代わりにLPC（線形予測分析）を用いることも可能である。また、映像入力部101から入力された映像の音データは、他方で音響検出部106に入力される。

【0016】図2は、本発明の一実施形態例のエッジ検出部102、音楽検出部103、音声検出部104の処理を示したフローチャートである。以下、図1及び図2を参照してそれらの動作例を説明する。

【0017】エッジ検出部102のFFT処理201によってスペクトログラムが生成される。その際のフレーム長は、数十～百ミリ秒で、検出区間は、数秒である。

【0018】図3に、生成されたスペクトログラムの様子を簡略化して示す。スペクトログラムは、実際には、

濃淡画像として得られる。301は、音楽成分のスペクトルの軌跡であり、302は、音声成分のスペクトルの軌跡である。音楽成分は、周波数方向に安定した軌跡を描くので、この性質を利用して検出する。まず、周波数*i*における時間方向のエッジE*D**i*をエッジ検出処理202で微分オペレータを用いて検出する。得られたエッジE*D**i*の値をエッジの閾値処理203で閾値TH1と比較し、エッジE*D**i*の値が閾値TH1よりも大きい場合には、音声検出の前処理として周波数*i*のスペクトルをエッジ消去、補間処理204において0にし、エッジを消去する。また、近傍のスペクトルの値を用いて消去されたスペクトルは、線形補間される。この処理を全ての帯域について繰り返す。繰り返し判定処理205において、*i*が*n*-1と等しくなれば繰り返しを終える。ここで*n*はFFTのフレーム長のポイント数である。

【0019】次に、エッジの強さの総和をエッジ強度算出処理206で算出し、エッジ強度の閾値処理207において、算出されたエッジの強さが閾値TH2よりも大きい場合に音楽が存在すると判断する。

【0020】図3の302に示すように、音声成分は時間的に変動する等間隔の縞模様として現れるので、エッジ強度算出処理206と平行してスペクトログラムにくし形フィルタ処理208を施し、フィルタ出力の閾値処理209において、フィルタ処理の出力が閾値TH3よりも大きければ音声が存在すると判断する。

【0021】図4は、本発明の一実施形態例の図1の音響検出部106の処理を示したフローチャートである。音響の種類例としては、笑声、歓声、拍手、雑踏、機械の音等が考えられる。ここでは、笑声、歓声、拍手を例に取って説明する。

【0022】笑声、歓声、拍手のような音響は、明確な構造がスペクトルに現れないため、ベクトル量子化を利用して検出する。まず、各々の音響データのサンプルを用意し、符号帳生成部105で符号帳を作成する。使用するベクトルの特徴量としては、数十～百ミリ秒のフレーム長で、16次元程度の線形予測係数を用いる。LPCケプストラム、FFTケプストラム、フィルタバンク出力等を用いることも可能である。サンプルデータは、多いほど良好な結果を得ることができる。笑声、歓声、拍手の3つのカテゴリーに分類するため、各サンプルデータの係数から3つ以上のクラスタを生成する。以下では、クラスタの数が3つの場合を例に取り説明する。まず、クラスタの重心ベクトルをC1、C2、C3とする。C1、C2、C3が、笑声、歓声、拍手のどの重心ベクトルに対応するかは、カテゴリーが既知のサンプルデータが最も近い重心ベクトルを調べることで、容易に分かる。

【0023】入力された映像の音データの線形予測係数は線形予測係数算出処理401で算出され、各々の重心ベクトルとの距離L*i*がベクトル距離算出処理402で

算出される。次に、最小距離ベクトルの閾値処理403において重心ベクトルとの距離 L_i の大きさを調べ、閾値TH4よりも大きい場合には、3つのカテゴリーには属しないと判断し、非音響と判断される。閾値TH4よりも小さい場合には、最小距離ベクトル判別処理404、最小距離ベクトル判別処理405により重心ベクトルとの距離 L_i の中で最も距離の短いものを選択し、対応するカテゴリーに属すると判断する。図4では、C1, C2, C3が各々、笑声、歓声、拍手に対応している場合を示している。

【0024】特徴音検出部102で検出された音の始点と終点の位置は、属性情報の一部として属性情報蓄積部107にタイムコードや、先頭からのバイト数等のフォーマットで記録される。

【0025】映像判別部106では、属性情報蓄積部107から情報を読み出し、映像シーケンス全体における各々の音の含有率を算出し、分類ベクトル $V(v_1, v_2, v_3, v_4, v_5, v_6)$ を求める。ここで、 $v_1, v_2, v_3, v_4, v_5, v_6$ は、各々、音楽、音声、笑声、歓声、拍手、音楽と音声重なっている区間、の含有率である。

【0026】分類ベクトルを用いて映像を分類する際には、音響検出と同様にベクトル量子化が用いられる。様々な映像サンプルを用いて分類ベクトルを求め、必要なジャンルの数だけクラスタリングを行い、重心ベクトルを求める。入力された映像の分類ベクトルと重心ベクトルの距離を算出し、最も近いクラスタに割り当てる。形成されるクラスタは、一般的に用いられるジャンルと必ずしも一致しないが、音声が多く、音楽が少なければニュースや教育、逆の場合は音楽、笑声が多い場合はコメディ等といった分類が可能である。

【0027】図5は、本実施形態例の映像分類装置をソフトウェアで実現した場合の処理を示すフローチャートである。映像は、まず、符号帳生成段階500で、音響の学習データから符号帳が生成され、映像入力段階501から入力され、エッジ検出段階502で周波数解析、エッジ検出が行われる。また、必要に応じてエッジの削除、補間が行われる。音楽検出段階503および音声検出段階504では、各々、エッジの強さ、くし形フィルタを用いて音楽および音声検出される。音響検出段階505では、ベクトル量子化を用いて、笑声、歓声、拍手が検出される。検出された音の始点と終点の情報は、属性情報検出段階506で蓄積され、映像シーケンスの最後に到達した時点で映像判別段階507において映像が分類される。

【0028】

【発明の効果】以上説明したように、本発明は以下のような効果を奏する。

【0029】(1) 映像情報に含まれる音情報から音楽、音声、笑声、歓声、拍手を検出し、検出された音情

報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置のパターンを比較するようにしたので、映像を広範囲なカテゴリーに分類することができる。

【0030】(2) スペクトログラムの一定周波数における時間方向のエッジの強さを検出するようにした場合には、特に音楽を容易に検出することができる。

【0031】(3) スペクトログラムのエッジの強い部分を除去した後に、くし形フィルタを用いてハーモニク構造を検出するようにした場合には、特に言葉などの音声も容易に検出することができる。

【0032】(4) 参照音として一種類の音響のみを含む音情報の特徴ベクトルと、該符号帳の重心との距離を算出し、距離が最も近くなる頻度の高い該符号帳の重心と、該映像情報に含まれる音情報の特徴ベクトルとの距離を検出の判定基準として用いるようにした場合には、特に学習した音響を容易に検出することができる。

【0033】(5) 検出された音情報の種類、各々の区間の長さ、種類毎の全体の長さ、各々の区間の位置を分類ベクトルとして符号帳を作成し、判別の判定基準に、該符号帳の重心と、該映像情報に含まれる音情報の分類ベクトルとの距離を用いるようにした場合には、特に映像を容易に広範囲なカテゴリーに分類することができる。

【図面の簡単な説明】

【図1】本発明の一実施形態例の映像分類装置の概略構成を示すブロック図である。

【図2】上記実施形態例の特徴音検出部分における音楽と音声の検出処理を示すフローチャートである。

【図3】上記実施形態例のエッジ検出部において得られたサウンドスペクトログラムの様子を示す概念図である。

【図4】上記実施形態例の特徴音検出部分における笑声、歓声および拍手の検出処理を示すフローチャートである。

【図5】上記実施形態例の映像分類装置を計算機を用いてソフトウェア的に実現した場合の処理の流れを示すフローチャートである。

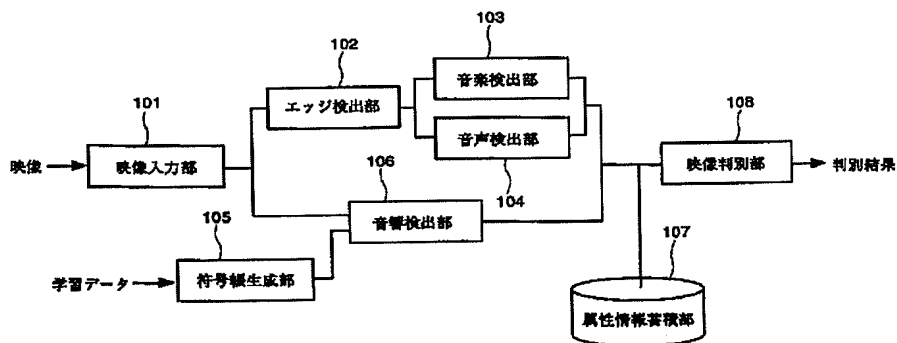
【符号の説明】

- 101…映像入力部
- 102…エッジ検出部
- 103…音楽検出部
- 104…音声検出部
- 105…符号帳生成部
- 106…音響検出部
- 107…映像判別部
- 108…属性情報蓄積部
- 201…FFT（高速フーリエ変換）処理
- 202…エッジ検出処理
- 203…エッジの閾値処理
- 204…エッジ消去、補間処理
- 205…繰り返し判定処理

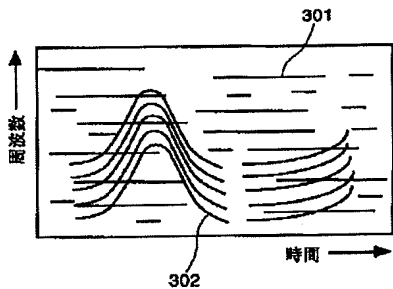
206…エッジ強度算出処理
 207…エッジ強度の閾値処理
 208…くし形フィルタ処理
 209…フィルタ出力の閾値処理
 301…音楽スペクトルピーク
 302…音声スペクトルピーク
 401…線形予測係数算出処理
 402…ベクトル距離算出処理
 403…最小距離ベクトルの閾値処理
 404…最小距離ベクトル判別処理

405…最小距離ベクトル判別処理
 500…符号帳生成段階
 501…映像入力段階
 502…エッジ検出段階
 503…音楽検出段階
 504…音声検出段階
 505…音響検出段階
 506…属性情報蓄積段階
 507…映像判別段階

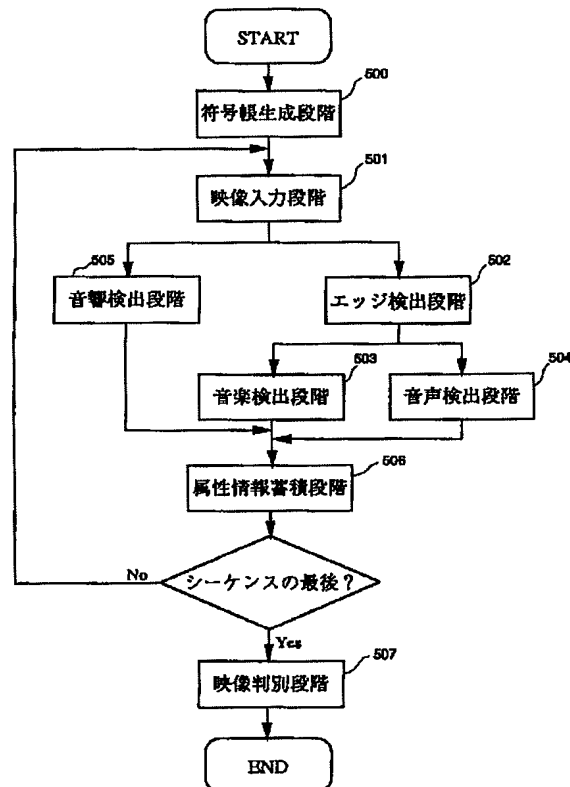
【図1】



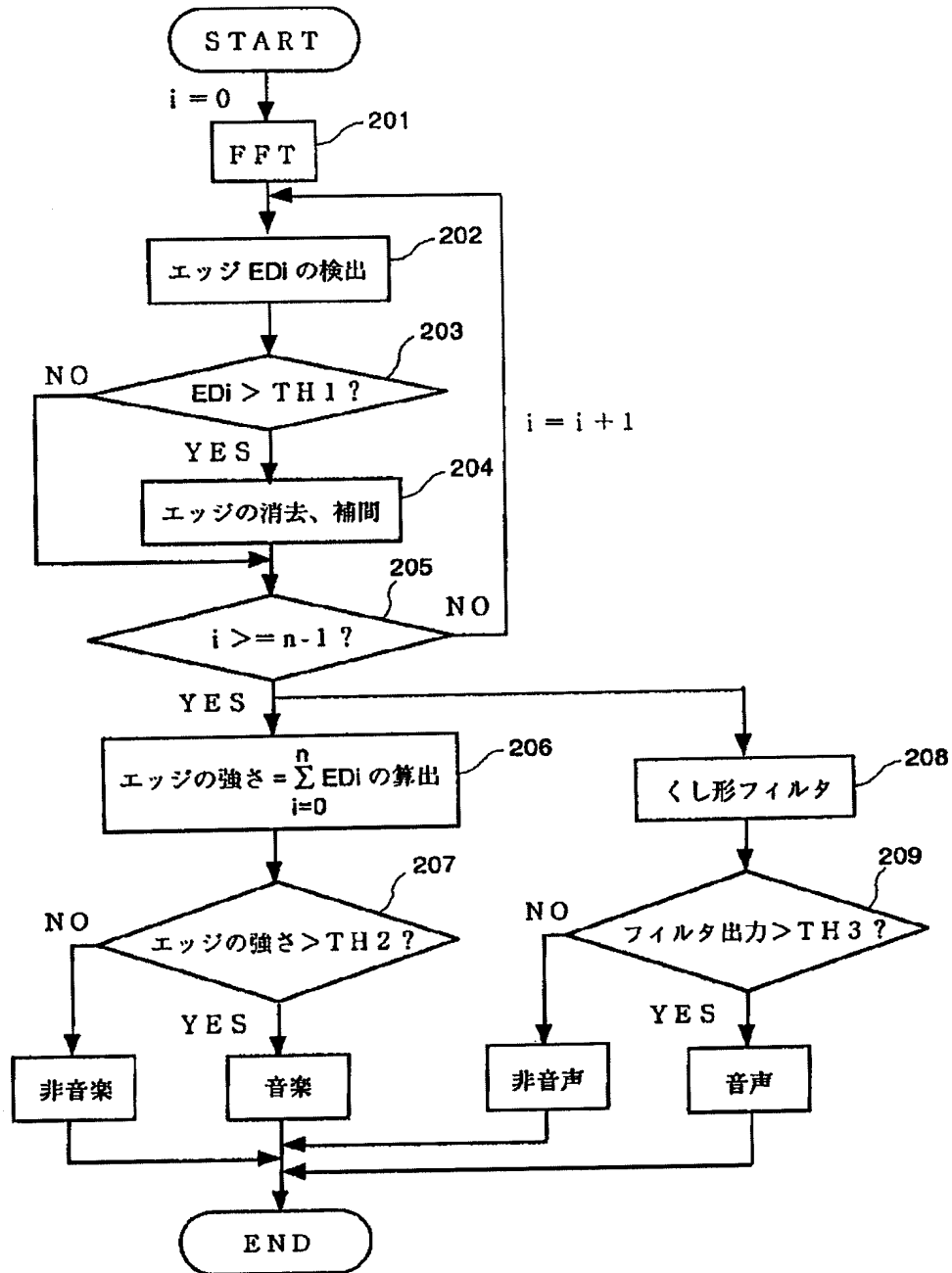
【図3】



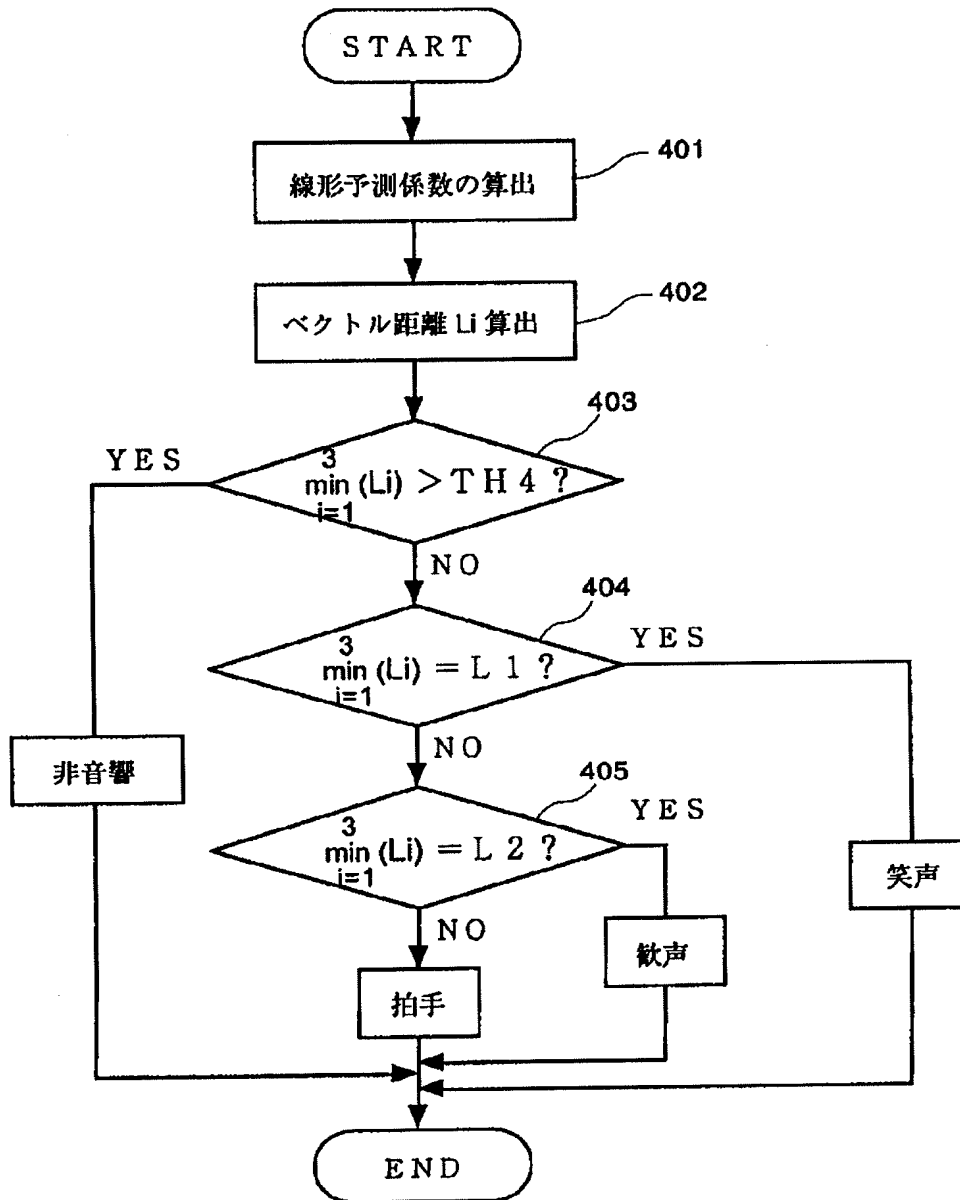
【図5】



【図2】



【図4】



PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-187182

(43)Date of publication of application : 14.07.1998

(51)Int.Cl. G10L 3/00
G10L 3/00
G10L 3/00
H04N 5/91

(21)Application number : 08-340293 (71)Applicant : NIPPON TELEGR & TELEPH
CORP <NTT>

(22)Date of filing : 20.12.1996 (72)Inventor : MINAMI KENICHI
AKUTSU AKITO
TONOMURA YOSHINOBU

(54) METHOD AND DEVICE FOR VIDEO CLASSIFICATION

(57)Abstract:

PROBLEM TO BE SOLVED: To provide the method and the device in which an analysis is conducted for the sound information included in video information and the video is classified into categories without being influenced by conventional categories.

SOLUTION: A music detecting section 103 conducts a frequency analysis of the sound information of the inputted video information, detects the stability of the spectra and detects a music. A voice detecting section 104 detects the harmonic structure of the spectrum and detects voices. On the other hand, a code table generating section 105 generates a code table. Then, an acoustic detecting section 106 compares the sound information of the inputted video information with the feature vectors of the code table and the kinds of acoustics are detected by the closeness of the distance between the sound information and the feature vectors. An attribute information accumulating section 107 records the position of the segment of the detected sound information and extracts the kinds of the detected sound information, the length of each segment, the total length of every kind and the pattern of the position of each segment and a video discriminating section 108 discriminates the kind of the video information.

CLAIMS

[Claim(s)]

[Claim 1] An image classifying method which detects musical sound and the section

when at least one of sound exists from sound information characterized by comprising the following which inputs video information and is included in this inputted video information and distinguishes a kind of image with occurrence patterns of the this detected section.

A video input stage of carrying out an A/D conversion and inputting digital video information when video information is an analog.

An edge detection stage of conducting frequency analysis of the sound information included in this video information and detecting the stability of a spectrum.

A music detection stage which detects music from the stability of this spectrum.

A voice detection stage of detecting harmonic structure of this spectrum and

detecting a sound A code book generation phase which vector-quantizes an acoustic feature vector as learned data and generates a code book A sound

detection stage which compares a generated this code book with a feature vector of sound information included in this video information and detects sound with a

near distance A stage according to image format which extracts one or more [of a pattern of an attribution information accumulation stage which records a position

of the section according to kind of detected this sound information and a kind of this detected sound information the length of each section length for every whole

kind and a position of each section] and distinguishes a kind of this video information.

[Claim 2] The image classifying method according to claim 1 characterized by what a differentiation operator of a frequency direction detects edge for in said edge detection stage from a spectrogram which arranged said spectrum in a time direction.

[Claim 3] The image classifying method according to claim 2 characterized by what music is detected for from strength of edge of a time direction in constant frequency of said spectrogram in said music detection stage.

[Claim 4] The image classifying method according to claim 2 or 3 characterized by what a radial fin type filter is used harmonic structure is detected and a sound is detected for after removing a portion with strong edge of said spectrogram in said voice detection stage.

[Claim 5] A feature vector of sound information which contains only one kind of sound as a reference sound in said sound detection stage An image classifying method given in either of claims 1, 2, 3 and 4 characterized by what distance with the center of gravity of said code book is computed and is used as a judging standard of detection of distance of the center of gravity of this code book with high frequency where distance becomes the nearest and a feature vector of sound information included in said video information.

[Claim 6] A stage according to said image format creates a code book by making a kind of detected sound information the length of each section length for every whole kind and a position of each section into a classification vector and The center of gravity of this code book An image classifying method given in either of claims

1234 and 5 characterized by what distance with a classification vector of sound information included in said video information is used for a distinction standard for. [Claim 7] An image classifying apparatus which detects musical sound and the section when at least one of sound exists from sound information characterized by comprising the following which inputs video information and is included in this inputted video information and distinguishes a kind of image with occurrence patterns of the this detected section.

A video input section which carries out an A/D conversion and inputs digital video information when video information is an analog.

An edge detection section which conducts frequency analysis of the sound information included in this video information and detects the stability of a spectrum.

A music primary detecting element which detects music from the stability of this spectrum.

A voice detection part which detects harmonic structure of this spectrum and detects a sound A code book generation part which vector-quantizes an acoustic feature vector as learned data and generates a code book A sound primary detecting element which compares a generated this code book with a feature vector of sound information included in this video information and detects sound with a near distance A part according to image format which extracts an attribution information accumulating part which records a position of the section according to detected this sound information a kind of this detected sound information the length of each section length for every whole kind and a pattern of a position of each section one or more and distinguishes a kind of this video information.

[Claim 8] The image classifying apparatus according to claim 7 characterized by what said edge detection section is what detects edge from a spectrogram which arranged said spectrum in a time direction with a differentiation operator of a frequency direction.

[Claim 9] The image classifying apparatus according to claim 8 characterized by what said music primary detecting element is what detects music from strength of edge of a time direction in constant frequency of said spectrogram.

[Claim 10] The image classifying apparatus according to claim 8 or 9 characterized by what is been what uses a radial fin type filter detects harmonic structure and detects a sound after said voice detection part removes a portion with strong edge of said spectrogram.

[Claim 11] A feature vector of sound information in which said sound primary detecting element contains only one kind of sound as a reference sound An image classifying apparatus given in either of claims 7 8 9 and 10 characterized by what is been what is used as a judging standard of detection of distance of the center of gravity of this code book with high frequency where distance with the center of gravity of said code book is computed and distance becomes the nearest and a feature vector of sound information included in said video information.

[Claim 12] A part according to said image format makes a classification vector a

kind of detected sound informationthe length of each sectionlength for every whole kindand a position of each sectioncreate a code bookand The center of gravity of this code bookAn image classifying apparatus given in either of claims 78910and 11 characterized by what is been what uses for a distinction standard distance with a classification vector of sound information included in said video information.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[Field of the Invention]In order to treat an image efficientlythe art which gives the attribution information of an image automatically is required. Attribution information is used for edit of an imageprocessinga classificationetc. in the related field of image work. This invention extracts the characteristic quantity contained in an imageand relates to the art of classifying an image according to characteristic quantity.

[0002]

[Description of the Prior Art]It is indispensable to divide roughly what kind of things the contents of the image arewhen treating efficiently a lot of images used by a system like video on demand. Although the image is mainly classified into newsa sporta dramaa moviemusicdocumentaryeducationvarietyanimeetc. nowthe method of identifying some automatically among these is proposed. In "S. Fischer et.al:Automatic Recognition of Film GenresACM Multimedia'95and pp.295-301." The change of a scene and a motion of a camera are detected from the sexual desire news of a pictureit combines with change of the amplitude of sound informationand the classification of newsa sport (tennis and car race)animeand commercials is performed. If there are few motions of a camera and there is a repetition (sound which hits the ball of tennis) of news and a periodic sounda sportIf the place where language broke off has few noises and the whole will become black at the change of anime (there are few background sounds because of postrecording)and a scenethe typical feature regarded as having called it commercials for every genre will be used.

[0003]

[Problem(s) to be Solved by the Invention]According to the above-mentioned conventional artthe image is mainly classified based on picture informationand detailed analysis about sound information is not conducted with it. Since the peculiar feature is restricted for every genre detectable from picture informationthe range which can be classified is narrow. The genre which cannot be classified exists in the top-down method that the feature for every genre defined from the former as mentioned above is found out.

[0004]On the other handthe sound information included in an image is reflecting the contents of the image welland tends to detect the feature peculiar to the kind

of contents. It is possible to realize the classifying method which took in the bottom-up element by detecting the characteristic sound which analyzes sound information and is looked at by the general image and classifying an image from the occurrence patterns.

[0005] The purpose of this invention analyzes the sound information included in video information and there is in providing the image classifying method and device which classify an image into the category which is not caught by the existing genre.

[0006]

[Means for Solving the Problem] In order to attain the above-mentioned purpose an image classifying method of this invention A video input stage of carrying out an A/D conversion and inputting digital video information when video information is an analog A music detection stage which conducts frequency analysis of the sound information included in this video information detects the stability of a spectrum and detects music A voice detection stage of detecting harmonic structure of this spectrum and detecting a sound A code book generation phase which vector-quantizes an acoustic feature vector as learned data and generates a code book A sound detection stage which compares a feature vector of a generated code book and sound information included in this video information and detects sound with a near distance An attribution information accumulation stage which records a position of the section according to kind of detected this sound information By having a stage according to image format which extracts a kind of detected this sound information the length of each section length for every whole kind and a pattern of a position of each section one or more and distinguishes a kind of this video information. It becomes possible to detect music a sound and the section when at least one of sound exists from sound information included in inputted video information to distinguish a kind of image and to classify into a wide range category according to occurrence patterns of the this detected section.

[0007] A video input section which carries out the A/D conversion of the image classifying apparatus of this invention when video information is an analog and inputs digital video information A music primary detecting element which conducts frequency analysis of the sound information included in this video information detects the stability of a spectrum and detects music A voice detection part which detects harmonic structure of this spectrum and detects a sound A code book generation part which vector-quantizes an acoustic feature vector as learned data and generates a code book A sound primary detecting element which compares a feature vector of a generated code book and sound information included in this video information and detects sound with a near distance An attribution information accumulating part which records a position of the section according to kind of detected this sound information By providing a part according to image format which extracts a kind of detected this sound information the length of each section length for every whole kind and a pattern of a position of each section one or more and distinguishes a kind of this video information. It becomes possible to detect music a sound and the section when at least one of sound exists from sound information included in inputted video information to distinguish a kind

of image and to classify into a wide range category according to occurrence patterns of the this detected section.

[0008]In above image classifying method and device detecting strength of edge of a time direction in constant frequency of a spectrogram enables it to detect music easily.

[0009]After removing a portion with strong edge of this spectrogram even when music has lapped it becomes possible by using a radial fin type filter and detecting harmonic structure to detect a sound easily.

[0010]A feature vector of sound information which contains only one kind of sound as a reference sound distance with the center of gravity of this code book is computed and distance becomes possible [detecting learned sound easily by using as a judging standard of detection of distance of the center of gravity of this code book with high frequency which becomes the nearest and a feature vector of sound information included in this video information].

[0011]Create a code book by making a kind of detected sound information the length of each section length for every whole kind and a position of each section into a classification vector and The center of gravity of this code book It becomes possible to classify an image according to using for a distinction standard distance with a classification vector of sound information included in this video information easily.

[0012]

[Embodiment of the Invention]Next an embodiment of the invention is described in detail with reference to drawings.

[0013]Drawing 1 is a block diagram showing the outline composition of the image classifying apparatus of the example of 1 embodiment of this invention.

[0014]The video input section 101 which carries out the A/D conversion of the image classifying apparatus of this example of an embodiment when video information is an analog and is inputted The edge detection section 102 which conducts frequency analysis of the sound information detects the edge of a sound spectrogram and is removed if needed The music primary detecting element 103 which detects music from sound information and the voice detection part 104 which detects a sound The code book generation part 105 which generates a code book from acoustic learned data and the learned sound and the sound primary detecting element 106 which detects the sound of an identical kind The parts 108 according to image format which distinguish the kind of video information are consisted of by the attribution information accumulating part 107 which records the position of the section of the detected sound the kind of detected sound information the length of each section the length for every whole kind and the position of each section.

[0015]The sound data of an image inputted from the video input section 101 is inputted into the edge detection section 102 by one side FFT (Fast Fourier Transform) processing is carried out by the edge detection section 102 and the sound spectrogram of the length for about several seconds is generated. Here it is also possible to use LPC (linear predictive coding) instead of FFT. The sound data

of an image inputted from the video input section 101 is inputted into the sound primary detecting element 106 on the other hand.

[0016] Drawing 2 is the flow chart which showed processing of the edge detection section 102 of the example of 1 embodiment of this invention the music primary detecting element 103 and the voice detection part 104. Hereafter those examples of operation are explained with reference to drawing 1 and drawing 2.

[0017] A spectrogram is generated by the FFT processing 201 of the edge detection section 102. The frame length in that case is tens-100 milliseconds and a detection interval is several seconds.

[0018] The situation of the generated spectrogram is simplified and shown in drawing 3. A spectrogram is actually obtained as a shade image. 301 is a locus of the spectrum of a music ingredient and 302 is a locus of the spectrum of a voice component. Since the locus stable in the frequency direction is drawn a music ingredient is detected using this character. First the edge ED_i of the time direction in the frequency *i* is detected using a differentiation operator by the edge detection process 202. When the value of the edge ED_i is larger than threshold TH1 in the value of the obtained edge ED_i at the threshold process 203 of edge as compared with threshold TH1 the spectrum of the frequency *i* is set to 0 in edge elimination and the interpolation processing 204 as pretreatment of voice detection and edge is eliminated. Linear interpolation of the spectrum eliminated using the value of a nearby spectrum is carried out. This processing is repeated about all the zones. In the repetition decision processing 205 if *i* becomes equal to *n*-1 a repetition will be finished. *n* is a point size of the frame length of FFT here.

[0019] Next total of the strength of edge is computed by the edge intensity calculation processing 206 and in the threshold process 207 of edge intensity when the strength of the computed edge is larger than threshold TH2 it is judged that music exists.

[0020] Since a voice component appears as a striped pattern at equal intervals changed in time as shown in 302 of drawing 3 in parallel with the edge intensity calculation processing 206 radial-fin-type-filter processing 208 is performed to a spectrogram and in the threshold process 209 of a filter output if the output of filtering is larger than threshold TH3 it will be judged that a sound exists.

[0021] Drawing 4 is the flow chart which showed processing of the sound primary detecting element 106 of drawing 1 of the example of 1 embodiment of this invention. As an example of an acoustic kind the sound of ****a cheer applause bustle and machinery etc. can be considered. Here it explains taking the case of ****a cheer and applause.

[0022] Since a clear structure does not appear in a spectrum ****a cheer and sound like applause are detected using vector quantization. First the sample of each sound data is prepared and a code book is created by the code book generation part 105. As characteristic quantity of the vector to be used it is the frame length of tens-100 milliseconds and the linear predictor coefficients of about 16-dimensional one are used. It is also possible to use LPC cepstrum FFT cepstrum a filter bank output etc. Sample data can obtain such a good result that it

is large. In order to classify into ***a cheer and three categories of applause the cluster of or more three ** is generated from the coefficient of each sample data. Below it explains taking the case of the case where the number of clusters is three. First the center-of-gravity vector of a cluster is set to C1C2 and C3. With ***a cheer and which center-of-gravity vector of applause C1C2 and C3 deal is that the sample data whose category is known investigates the nearest center-of-gravity vector and it is known easily.

[0023] The linear predictor coefficients of the inputted sound data of an image are computed by the linear-predictor-coefficients calculation processing 401 and the distance L_i with each center-of-gravity vector is computed by the vector distance calculation processing 402. Next in the threshold process 403 of a shortest distance vector the size of the distance L_i with a center-of-gravity vector is investigated and when larger than threshold TH4 it judges that it does not belong to three categories and is judged as non-sound. In being smaller than threshold TH4 it judges that what has the shortest distance is chosen in the distance L_i with a center-of-gravity vector by the shortest distance vector discrimination processing 404 and the shortest distance vector discrimination processing 405 and it belongs to a corresponding category. Drawing 4 shows the case where C1C2 and C3 support ***a cheer and applause respectively.

[0024] The position of the starting point of a sound and a terminal point detected in the feature sound primary detecting element 102 is recorded on the attribution information accumulating part 107 in the format of a time code the number of bytes from a head etc. as a part of attribution information.

[0025] In the part 106 according to image format information is read from the attribution information accumulating part 107 the content of each sound in the whole picture sequence is computed and the classification vector V ($v_1v_2v_3v_4v_5v_6$) is searched for. Here $v_1v_2v_3v_4v_5$ and v_6 are the content of section ** to which the sound has lapped with musica sound ***a cheer applause and music respectively.

[0026] When classifying an image using a classification vector vector quantization as well as sound detection is used. A classification vector is searched for using various image samples only the number of required genres clusters and a center-of-gravity vector is searched for. The inputted distance of the classification vector of an image and a center-of-gravity vector is computed and it assigns the nearest cluster. Although the cluster formed is not necessarily in agreement with the genre generally used if there is little music when reverse and there are [news education and] much music and ***the classification of a comedy etc. is possible [there are many sounds and].

[0027] Drawing 5 is a flow chart which shows processing when software realizes the image classifying apparatus of this example of an embodiment. An image is the code book generation phase 500 first a code book is generated from acoustic learned data it is inputted from the video input stage 501 and frequency analysis and edge detection are performed in the edge detection stage 502. Deletion of edge and interpolation are performed if needed. In the music detection stage 503

and the voice detection stage 504music and a sound are respectively detected using the strength of edgeand a radial fin type filter. In the sound detection stage 505****a cheerand applause are detected using vector quantization. The detected information on the starting point of a sound and a terminal point is accumulated by the attribution information detection stage 506and when the last of a picture sequence is reachedan image is classified in the stage 507 according to image format.

[0028]

[Effect of the Invention]As explained abovethis invention does the following effects so.

[0029](1) Since musica sound****a cheerand applause are detected from the sound information included in video information and it was made to compare the kind of detected sound informationthe length of each sectionthe length for every whole kindand the pattern of the position of each sectionan image can be classified into a wide range category.

[0030](2) Especially when the strength of the edge of the time direction in the constant frequency of a spectrogram is detectedmusic can be detected easily.

[0031](3) Especially when a radial fin type filter is used and harmonic structure is detected after removing a portion with strong edge of a spectrogramsoundsuch as languagecan be detected easily.

[0032](4) The feature vector of the sound information which contains only one kind of sound as a reference soundDistance with the center of gravity of this code book is computedand especially when it is made to use as a judging standard of detection of the distance of the center of gravity of this code book with high frequency where distance becomes the nearestand the feature vector of the sound information included in this video informationthe learned sound can be detected easily.

[0033]Create a code book by making the kind of detected sound informationthe length of each sectionthe length for every whole kindand the position of each section into a classification vectorand to the judging standard of distinction (5) The center of gravity of this code bookEspecially when distance with the classification vector of the sound information included in this video information is usedan image can be easily classified into a wide range category.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1]It is a block diagram showing the outline composition of the image classifying apparatus of the example of 1 embodiment of this invention.

[Drawing 2]It is a flow chart which shows the detection processing of the music in the feature sound detection section of the above-mentioned example of an embodimentand a sound.

[Drawing 3]It is a key map showing the situation of the sound spectrogram

obtained in the edge detection section of the above-mentioned example of an embodiment.

[Drawing 4] It is a flow chart which shows **** in the feature sound detection section of the above-mentioned example of an embodiment and the detection processing of applause.

[Drawing 5] It is a flow chart which shows the flow of processing at the time of realizing the image classifying apparatus of the above-mentioned example of an embodiment by software using a computer.

[Description of Notations]

- 101 -- Video input section
 - 102 -- Edge detection section
 - 103 -- Music primary detecting element
 - 104 -- Voice detection part
 - 105 -- Code book generation part
 - 106 -- Sound primary detecting element
 - 107 -- Part according to image format
 - 108 -- Attribution information accumulating part
 - 201 -- FFT (Fast Fourier Transform) processing
 - 202 -- Edge detection process
 - 203 -- Threshold process of edge
 - 204 -- Edge elimination/interpolation processing
 - 205 -- Repetition decision processing
 - 206 -- Edge intensity calculation processing
 - 207 -- Threshold process of edge intensity
 - 208 -- Radial-fin-type-filter processing
 - 209 -- Threshold process of a filter output
 - 301 -- Music spectral peak
 - 302 -- Voice spectral peak
 - 401 -- Linear-predictor-coefficients calculation processing
 - 402 -- Vector distance calculation processing
 - 403 -- Threshold process of a shortest distance vector
 - 404 -- Shortest distance vector discrimination processing
 - 405 -- Shortest distance vector discrimination processing
 - 500 -- Code book generation phase
 - 501 -- Video input stage
 - 502 -- Edge detection stage
 - 503 -- Music detection stage
 - 504 -- Voice detection stage
 - 505 -- Sound detection stage
 - 506 -- Attribution information accumulation stage
 - 507 -- Stage according to image format
-